MPEG-4, ИЛИ СТРУКТУРИЗАЦИЯ МУЛЬТИМЕДИА (часть 1)

Статья была опубликована в журнале PC Magazine, дополнена и переработана специально для журнала РЭТ.

Сергей Новосельцев (Москва)

Формирующийся на наших глазах цифровой мир основан на новых информационных технологиях. Формат VHS вытесняется форматом DVD, компакт-кассета уступила место компакт-диску с записью в формате MP3, пройдет немного времени, и аналоговое телевидение заменит цифровой формат DTV. Стандарт MPEG-4, объединяя все общепринятые форматы, содержит потрясающие возможности для формирования медиасреды и, скорее всего, будет определять ее развитие в ближайшее десятилетие. Статья адресована читателям, желающим заглянуть в медиамир недалекого будущего.

ВНИМАНИЕ: MPEG-4!

Стандарты серии MPEG для сжатия видео-/аудиоданных разрабатывает группа Motion Picture Expert Group Международной организации по стандартам (ISO). Напомним, что MPEG-1 (принят в качестве официального стандарта в 1992 г.) разрабатывался для записи видео на CD-ROM (скорость около 1,5 Мбит/с) и получил широкое распространение во многом благодаря дискам VideoCD (до сих пор очень популярным в Азии, в частности в Китае). MPEG-2 (1994 г.) предназначен для работы с видео вещательного качества (скорость потока данных 3...15 Мбит/с) и сегодня повсеместно используется в цифровом телевещании, а также при кодировании видеоматериалов для дисков DVD-Video. Группа MPEG начала было работу над стандартом MPEG-3, который должен был обеспечить унификацию в компрессии потоков данных со скоростями 20...40 Мбит/с для телевидения высокой четкости (HDTV), но довольно быстро обнаружилось, что принципиальной разницы в подходах между MPEG-2 и MPEG-3 нет, в результате чего разработка последнего была прекращена, а рамки MPEG-2 расширены. MPEG-4 же, подобно Золушке, из стандарта «мультимедиа для бедных» с дергающейся картинкой в маленьком окошке превращается в главное действующее лицо мира мультимедиа (фактически подчинив себе и «старших сестер», области применения которых теперь можно трактовать как частные случаи – как способы кодирования одного из многочисленных типов данных, предусмотренных MPEG-4). Важность этого стандарта трудно переоценить, он гораздо больше, чем просто описание правил кодирования. По существу, он претендует на то, чтобы стать – спустя полтора десятилетия после зарождения цифрового мультимедиа – единым концептуальным способом описания, представления и обработки мультимедиаданных на следующие десять лет.

К сожалению, у нас в стране состояние дел с MPEG-4 пока не слишком известно даже большинству тех специалистов, которые будут прямо вовлечены в его внедрение и применение, – а с ним придется иметь дело и программистам, и разработчикам декодеров, и создателям интерактивных мультимедиапродуктов, и телевизионным авторам и вещателям, и сетевым провайдерам, а затем и всем остальным, простым домашним потребителям. При этом, в отличие от MPEG-1 и -2, которые предстают некими строго локализованными «черными ящиками», сжимающими/разжимающими видео незаметно для пользователя, MPEG-4 будет повсюду и прямо повлияет на способ работы и мышления многих категорий специалистов, а бесчисленные заложенные в него потенциальные возможности придется внимательно изучать, чтобы опередить конкурентов.

MPEG-4 признан в качестве официального международного стандарта – ISO/IEC 14496 – в 1999 г., работа над его усовершенствованием продолжается, уже недалеко до массового выхода на рынок различных продуктов, построенных на его основе. Поэтому пришла пора поговорить о нем более подробно – чтобы не быть застигнутыми врасплох.

ОБЪЕКТНО-ОРИЕНТИРОВАННОЕ МУЛЬТИМЕДИА

Новое предназначение стандарта MPEG-4 в рабочих документах группы MPEG формулируется так: он задает принципы работы с контентом (цифровым представлением медиаданных) для трех областей: собственно интерактивного мультимедиа (включая продукты, распространяемые на оптических дисках и через Сеть), графических приложений (синтетического контента) и цифрового телевидения – DTV. При этом его главное достоинство, на мой взгляд, состоит в том, что он не просто оформляет ту или иную сложившуюся практику в качестве стандарта, но, подобно американской конституции, является опережающим, структурообразующим и фундаментальным законом, создающим основу для производства, распространения контента и способов доступа к нему в новой единой цифровой среде и открывающим – в случае его признания перечисленными отраслями – множество принципиально новых возможностей для авторов, дистрибьюторов и потребителей этого контента.

MPEG-4 – не только стандарт, фактически он задает правила организации среды, причем среды объектно-ориентированной. Он имеет дело не просто с потоками и массивами медиаданных, а с медиаобъектами – это ключевое понятие стандарта. Объекты могут быть аудио, видео, аудиовизуальными, графическими (плоскими и трехмерными), текстовыми. Они могут быть как «естественными» (записанными, отснятыми, отсканированными и т.п.), так и синтетическими (т.е. искусственно сгенерированными). Примерами объектов могут служить неподвижный фон, видеоперсонажи отдельно от фона (на прозрачном фоне), синтезированная на основе текста речь, музыкальные фрагменты, трехмерная модель, которую можно двигать и вращать в кадре, анимированный спрайт (о спрайтах см. в главе «Кодирование видео»). Медиаобъекты могут быть потоковыми. Каждый медиаобъект имеет связанный с ним набор дескрипторов, где и задаются все его свойства,

операции, необходимые для декодирования ассоциированных с ним потоковых данных, размещения в сцене, а также поведение и допустимые реакции на воздействия пользователя. Из объектов строятся сцены. Сцена имеет свою систему координат, в соответствии с которой размещаются объекты. Звуковые объекты также могут иметь (и менять во времени) координаты в пространстве сцены, благодаря чему достигаются стерео- и «окружающие» (surround) эффекты. Объекты могут быть элементарными (primitive) и составными (compound) – представляющими ту или иную композицию элементарных объектов (например, сгенерированный трехмерный телевизор, наложенная на его экран живая видеотрансляция и исходящий из его динамиков звук). Стандарт задает правила кодирования различных объектов, их иерархии и способы композиции при построении сцены, а также методы взаимодействия пользователя с отдельными объектами внутри сцены. Каждый объект имеет свою локальную систему координат – с ее помощью объект управляется в пространстве и во времени. При помещении объекта в сцену происходит преобразование его локальной системы координат в систему координат старшего по иерархии объекта или глобальную систему координат сцены. Объекты и сцена могут обладать поведением, контролируемым уровнем композиции при визуализации сцены (характер звука, цвет объекта и т.п.). Сцена описывается с помощью иерархической структуры; узлами этой структуры являются объекты, и она динамически перестраивается по мере того, как узлы-объекты добавляются, удаляются или заменяются.

В MPEG-4 определен двоичный язык описания объектов, классов объектов и сцен BIFS, который характеризуют как «расширение Си++». С помощью команд BIFS можно анимировать объекты, менять их координаты, размеры, свойства, задавать поведение, реакции на воздействия пользователя, менять свойства среды, изменять и обновлять сцену, выполнять 2D-геометрические построения и т.п. Поскольку язык двоичный, он весьма компактен и быстр в интерпретации. Согласно заявлениям разработчиков, многие концепции BIFS позаимствованы у VRML, и MPEG и Web 3D Consortium продолжают работу по сближению MPEG-4 и VRML.

АКТИВНАЯ ЗРИТЕЛЬСКАЯ ПОЗИЦИЯ

Для понимания революционной сущности MPEG-4 очень важно подчеркнуть, что окончательная сборка сцены (причем с возможностью добавления разного рода геометрических преобразований, визуальных и акустических эффектов реального времени), вообще говоря, происходит на приемном конце – в компьютере, приставке или телевизоре пользователя. Это, в частности, позволит в корне изменить (еще раз оговоримся – только после реального признания стандарта производителями телепрограмм и вещательными корпорациями и появления MPEG-4-совместимых приставок или телевизоров) всю концепцию современного телевидения. Каждый сам, наверное, представит количество степеней свободы, которое может получить телезритель. Вместо сегодняшнего плоского окошка, отображающего аудиовизуальный поток, где-то кем-то подготовленный и директивно выдаваемый в эфир, окошка, оставляющего только одну степень свободы – переключить канал («Я тебе покручу!») или вообще выключить телевизор, зритель получает некое подобие виртуального пространства, с которым он может взаимодействовать и которое (при соответствующей доброй воле производителя телепрограммы) может выстраивать удобным для себя образом. Простейший пример такого взаимодействия с контентом – динамический выбор той или иной камеры или повтора при просмотре спортивных передач (естественно, при многокамерных трансляциях по цифровым каналам – но это уже близко к реальности даже для России): фактически зритель – «Сам себе режиссер трансляции», его функции подобны тем, которые лет тридцать, сидя в телевизионном автобусе возле «Лужников» или «Динамо», бессменно выполняли для него Ян Садеков и Раиса Панина.

Но это – только цветочки. Среди допустимых в принципе пользовательских команд взаимодействия с контентом – изменение точки наблюдения, удаление, добавление и перемещение объектов внутри сцены, выбор той или иной языковой дорожки, активизация более или менее сложной цепочки событий путем «щелчка» на объекте, ввод команд с клавиатуры и т.п. Естественно, эти воздействия должны быть предусмотрены и разрешены создателями того или иного контента – в противном случае пользователь остается пассивным зрителем, наблюдающим сцены, построенные автором, режиссером (это должно развеять опасения некоторых авторов и вещателей относительно того, что с введением MPEG-4 они утратят возможность контролировать качество продукта, картинку, которую увидит зритель на своем экране, и в конечном итоге – эстетическое и эмоциональное воздействие произведения). Для отслеживания действий пользователя и описания реакций на них реализована структура событий из VRML. Опираясь на эту модель, авторы контента могут создавать действительно интерактивные произведения и передачи.

Добавим, что стандарт предусматривает как локальную отработку воздействий и команд пользователя в декодере (client side interaction), так и пересылку их для исполнения на передающую сторону по обратному «восходящему» каналу, если декодер обладает такой возможностью, а серверная сторона готова реагировать на запросы снизу (server side interaction).

КОДИРОВАНИЕ ВИДЕО

Как уже упоминалось, MPEG-4 начинал разрабатываться как способ передачи потоковых медиаданных, в первую очередь видео, по каналам с низкой пропускной способностью (4,8...64 Кбит/с), в том числе беспроводным. Сейчас эта часть представлена блоком VLBV Core (Very Low Bit-rate Video) – ядром, обеспечивающим работу с «видео, имеющим очень низкую скорость потока данных». Естественно, такое видео имеет ухудшенные характеристики как по разрешению (до так называемого разрешения CIF, Common Interchange Format, – 320 × 240), так и по частоте кадров (до 15 кадр/с); впрочем, прогресс методов сжатия постоянно повышает верхнюю границу характеристик – всего два года назад речь шла лишь о 176 × 144... Помимо эффективных и помехоустойчивых методов кодирования последовательностей подобных кадров, VLBV содержит предложения по реализации операций произвольного доступа к кадрам видеопоследовательности, а также быстрой «подмотки» видеоряда вперед

и назад. Это требуется, например, в бурно развивающейся области управления медиа–активами (Digital Asset Management) – для работы с базами видеоданных, хранящими видеоматериалы в низком разрешении (для целей быстрого поиска и оценки) и ссылки на места хранения соответствующих материалов в полном вещательном качестве.

Второй блок, отвечающий за работу с видео с большой скоростью потока, вплоть до вещательного качества по стандарту ITU–R 601, обеспечивает, в общем, те же функции, что и VLBV, однако здесь предусмотрены возможности работы с видео, имеющим не только прогрессивную, но и чересстрочную телевизионную развертку. Два названных блока обрабатывают обычные видеопотоки с прямоугольными кадрами и фактически включают в себя функциональность MPEG–1 и MPEG–2, а также кодирование «живых» текстур.

Особенно интересен третий блок – так называемые функции, зависящие от контента. Сюда входит обработка видео с произвольным силуэтом (с помощью 8–бит–механизма прозрачности или двоичных масок) для отдельного кодирования видеообъектов (например, «вырезанного» силуэта диктора) и интерактивных манипуляций с ними. Помимо обычных методов межкадрового кодирования – предсказания и компенсации движения, – предусмотрены механизмы работы со «спрайтами» – неподвижными изображениями, которые передаются в декодер лишь однажды и всякий раз подставляются в нужное место кадра из специального спрайтового буфера. Механизм спрайтов позволяет значительно снизить объем передаваемых данных и обеспечивает большую гибкость в построении сцен. Например, можно запускать различные объекты–спрайты (самолеты, автомобили) поверх «живого» видеофона или же построить (выделить из реальных съемок или сгенерировать) неподвижную спрайт–панораму шириной в несколько кадров для «задника» сцены (спортивная площадка и трибуны) и, запустив поверх нее «живые» видеообъекты (игроков), панорамировать камерой вправо–влево – в этом случае для каждого кадра достаточно передавать вместо полной картинки фона только параметры камеры – направление и наплыв (zoom). Для улучшения времени реакции спрайт–панорамы могут подкачиваться с «прогрессивным разрешением» – т.е. с постепенным улучшением разрешения, как картинки в Интернете.

Этот же блок отвечает за масштабируемость видеообъектов. Под этим термином подразумевается, что объекты кодируются таким образом, чтобы декодер имел возможность в случае ограничений пропускной способности сети или параметров самого декодера (недостаточная вычислительная мощность, малое разрешение дисплея) огрублять изображение, декодируя и выводя лишь часть передаваемой потоковой информации (например, уменьшая частоту или разрешение кадров, увеличивая зернистость), но сохраняя, тем не менее, адекватность передачи контента. Для видеопотоков предусмотрено до трех уровней зернистости. При кодировании неподвижных изображений и текстур в MPEG–4 применяется очень эффективный wavelet–алгоритм, обеспечивающий кодирование объектов произвольной формы, 11 уровней масштабируемости по разрешению и плавную масштабируемость по качеству картинки. Результирующий закодированный поток представляет собой «пирамиду» различных разрешений, и в приемнике картинка со временем «проявляется», улучшаясь настолько, насколько позволяет данная передающая среда.

СИНТЕТИЧЕСКИЕ ОБЪЕКТЫ И ЛИЦА

В MPEG–4 предусмотрены инструменты и алгоритмы для работы не только с видеообъектами, но и с объектами синтетическими, т.е. сгенерированными средствами компьютерной графики: каркасными представлениями (mesh) двух– и трехмерных моделей, потоками геометрических данных для анимирования этих моделей, с натуральными («живыми») или анимированными текстурами, которые могут на эти модели накладываться и т.п. Подобные объекты позволяют значительно сократить объем передаваемых данных, так как для их анимации бывает достаточно передать всего несколько параметров – все остальное будет сделано в декодере.

Среди синтетических объектов выделена в отдельный класс анимация человеческих лиц и фигур. В MPEG–4 установлены наборы управляющих параметров для задания особенностей лица (FDP), для его анимации (FAP) и интерполяции, контрольные точки в полигональной сетке, отвечающие за те или иные эмоции или движения (с весовыми коэффициентами) и т.п. Необходимые средства управления анимацией входят в язык BIFS. Лицо может быть сгенерировано в декодере на базе имеющейся в нем обобщенной модели и затем «индивидуализировано» с помощью FDP, либо желаемая конкретная модель (например, полученный с помощью трехмерного сканера «автопортрет») может быть загружена во входящем потоке. Мало этого, на построенную модель лица можно «натянуть» фото– или видеотекстуру лица конкретного человека, а затем «заставить» его произносить написанный текст. Средства синтеза речи на базе текстов (text–to–speech), предусмотренные в MPEG–4, не только генерируют необходимые фонемы, но могут также создавать поток данных для соответствующей анимации модели лица говорящего. Таким образом можно построить виртуального диктора, изображение удаленного абонента при «разговоре» в чате или отправить сетевым партнерам собственного аватара–дубля.

Имеются развитые средства работы с двумерными полигональными моделями, адаптации их под имеющийся видеоконтент для последующей анимации, – например, искажения текстур в соответствии с деформацией подложенной сетки и др. Использование этих средств позволяет выполнять многие функции: представление контуров объектов с помощью вершин сетки (вместо битовых масок), замещение в сцене «живых» видеообъектов синтетическими и т.д., – отсылаем читателя к описанию стандарта.

Сюда примыкают и средства учета точки наблюдения, которые работают как на клиентской, так и на серверной стороне (если имеется обратный канал): при наличии в трехмерном пространстве сцены объектов переднего плана те фрагменты изображения, которые заслонены для наблюдателя этими объектами, не передаются.

ЗВУК

Несмотря на отсутствие в названии группы MPEG даже намека на звук, ее эксперты весьма успешно

работают в этой области, и их предложения действительно становятся общеупотребительными стандартами, порой опережая разработки «профильных» звуковых организаций и фирм. При этом звуковая часть стандартов MPEG достаточно слабо связана с видеочастью, новые версии и алгоритмы, выбранные экспертами, просто добавляются к уже имеющимся функциям. Так, в частности, был добавлен (к ранее стандартизованным Уровням 1 и 2) формат сжатия MPEG Audio Уровень 3 для стандартов MPEG-1 и -2, разработанный специалистами Fraunhofer Institute for Integrated Circuits (IIS-A) и University of Erlangen в рамках проекта цифрового аудиовещания DAB. Этот стандарт сегодня известен всем под именем MP3 (не путать с MPEG-3). Он зажил самостоятельной, отдельной от видеоряда жизнью и грозит перевернуть весь бизнес звукозаписи из-за высокого качества, компактности сжатых им материалов и расцвета несанкционированного распространения их через Сеть. Его последователь, формат MPEG-2 AAC (Advansed Audio Coding), также разработанный в IIS-A (www.iis.fhg.de), соперничает с Dolby AC-3 в качестве многоканального формата записи звука для дисков DVD-Video. Этот формат обеспечивает по сравнению с MP3 еще более высокое качество звучания, лучшую степень сжатия и возможность работы с различными потоками, от моно- до многоканальных.

При всем множестве новаторских подходов MPEG-4, звуковые разделы стандарта – возможно, наиболее интересная и революционная его часть. Объектный подход к изображениям – откровение для телевидения, но в ряде систем анимации, в VRML он применялся и ранее. Что же касается объектного звука, то системы, сопоставимой с MPEG-4 по комплексности подхода, спектру примененных технологий и диапазону применений, просто не удается вспомнить. Она заслуживает отдельного разговора, а здесь мы можем лишь бегло перечислить ее возможности.

Как и другие типы объектов, аудиообъекты входят в структуру дерева сцены и описываются на языке BIFS, что позволяет располагать источники звука в трехмерном пространстве сцены, управлять их характеристиками и применять к ним различные эффекты независимо друг от друга, перемещать источник звука при перемещении связанного с ним визуального объекта и т.п. В следующей версии в стандарт будет добавлена возможность задания акустических параметров среды. Отметим, что все эффекты и анимации выполняются в декодере по командам, полученным во входном потоке, что уменьшает объем передаваемых данных и увеличивает гибкость.

Для кодирования аудиообъектов MPEG-4 предлагает наборы инструментов как для живых звуков, так и для синтезированных. MPEG-4 устанавливает синтаксис двоичных потоков и процесс декодирования в терминах наборов инструментов, это позволяет применять различные алгоритмы сжатия. Диапазон предлагаемых стандартом скоростей потока для кодирования живых звуков от 2 до 128 Кбит/с и выше. При кодировании с переменным потоком минимальная средняя скорость может оказаться еще меньше, порядка 1,2 Кбит/с. Для звука высшего качества применяется алгоритм AAC, который дает качество лучше, чем у CD, при потоке в 10 с лишним раз меньше. Другой возможный алгоритм кодирования живого звука – TwinVQ. Для кодирования речи предлагаются алгоритмы: HVXC (Harmonic Vector eXcitation Coding) – для скоростей потока 2...4 Кбит/с – и CELP (Code Excited Linear Predictive) – для скоростей 4...24 Кбит/с. Предусмотрены различные мехнизмы масштабируемости.

Особый раздел – синтез речи. На входы синтезатора поступает текст, а также различные параметры «окраски» голоса – ударения, изменения высоты тона, скорости произнесения фонем и т.п. Можно также задать для «говорящего» пол, возраст, акцент и т.п. В текст можно вставлять управляющую информацию, обнаружив которую, синтезатор синхронно с произнесением соответствующей фонемы передаст те или иные параметры или команды другим компонентам системы. Параллельно с голосом может генерироваться поток параметров для анимации лица. Отметим, что, как и всегда, MPEG-4 задает правила работы, интерфейс синтезатора, но не его внутреннее устройство.

Наконец, самая интересная часть звуковой составляющей – средства синтеза произвольных звуков и музыки. Здесь MPEG-4 предлагает в качестве стандарта подход, разработанный в колыбели многих передовых технологий – MIT Media Lab – и названный Structured Audio (SA) – «Структурированный звук». Опять-таки, это не конкретный метод синтеза, а формат описания методов синтеза, в котором можно описать любой из существующих методов (а также, как утверждается, будущих). Для этого вводятся два языка: SAOL (Structured Audio Orchestra Language) и SASL (Structured Audio Score Language). Как следует из названия, первый задает оркестр, а второй – то, что этот оркестр должен играть. Оркестр состоит из инструментов. Каждый инструмент представлен сетью элементов цифровой обработки сигналов – синтезаторов, цифровых фильтров, которые все вместе и синтезируют нужный звук. С помощью SAOL можно запрограммировать практически любой нужный инструмент, природный или искусственный звук. Сначала в декодер загружается набор инструментов, а затем поток данных SASL заставляет этот оркестр играть, управляя процессом синтеза. Таким образом обеспечивается одинаковое звучание на всех декодерах при очень низком входном потоке и высокой точности управления.

Стандартом допускается также управление, основанное на протоколе MIDI, – но этот метод не столь точен, а набор инструментов ограничен. Для простых декодеров стандартизован также формат для работы с волновыми таблицами (wavetable bank format) – в этом случае в декодер загружаются набор сэмплов и необходимые фильтры и эффекты.

Продолжение следует.

Литература

1. ISO/IEC JTC1/SC29/WG11 No. 2459 Overview of the MPEG-4 Standard. October 1998/Atlantic City.

2. ISO/IEC JTC1/SC29/WG11 No. 4668 Overview of the MPEG-4 Standard. March 2002.

3. Koenen R. MPEG-4. Multimedia For Our Time. IEEE Spectrum, February, 1999.

4. Биркмайер К. Общество Плоской Земли. «Мультимедиа. Цифровое видео», №4, 1998.

5. Фоминов О. Мультимедиа и сети. «Мультимедиа. Цифровое видео», №5-6, 7, 8, 1997.